National Statistical Service

Confidentiality Information Sheet 4

Confidentiality is about protecting the identity of individuals or organisations.

Confidentiality

How to confidentialise data: the basic principles

Managing the risks of identification in disseminated data (also called **disclosure control**) involves taking steps to evaluate and mitigate the risk that the identity of a particular person or organisation may be disclosed.

Risks of identification can be managed by confidentialising data. The aim is to protect the identity of a person or an organisation, while at the same time maximising the usefulness of the data.

In simple cases, data can be manually confidentialised. However, the use of software is sometimes necessary. Specialised skills and knowledge of the data are also required to correctly confidentialise a dataset to minimise the risk of identification.

This information sheet outlines some common techniques for confidentialising data. The information is provided as a guide only and gives simple examples, using data presented in tables, to illustrate the concepts. However, most of the techniques apply to both aggregate and microdata.

Issues specific to the management of the confidentiality of microdata are discussed in *Information Sheet 5*, '*Managing the risk of disclosure in the release of microdata*'.

Identification risks in aggregate data

Aggregate data can be disseminated in various forms including tables, maps and graphs. Aggregate data in any form can present an identification risk if individual responses can be estimated or derived from the output (for example, outliers in a graph).

Tables are the most common way of presenting aggregate data and are the focus of the following discussion.

There are two main types of tables:

- Count (frequency) tables: cells contain the number of individuals or organisations contributing to the cell (e.g. the number of people in various age groups, or the number of businesses in each industry).
- Magnitude tables: cells contain values calculated from a numeric response (e.g. total income or profit).

It is sometimes possible to deduce information about a particular person or organisation from these two types of tables.

For example, when a cell in a frequency table has a low *count* (that is, only a very small number of contributors), it may be possible to deduce information about a particular person or organisation using the information you already know and the additional information presented in the table. This poses an identification (or disclosure) risk.

In magnitude tables, table cells present an identification (or disclosure) risk when they are dominated by *values* relating to one or two businesses or individuals.

Further identification risks exist if users have access to multiple tables that contain some common elements. It may be possible to use information from one table to determine the identity of a person or organisation contributing to a second table. This means it is important to keep track of all information that is released from the dataset.

When should a cell be confidentialised?

To identify table cells that pose an identification (or disclosure) risk, **confidentiality rules** must be determined and applied to each cell in a table. If a cell fails such rules, then further investigation or action is needed to minimise the risk of identification.

Two common rules used to assess identification risk in a table cell are:

- The **frequency rule** (also called the threshold rule), which sets a threshold value for the minimum number of units (contributors) in any cell. Common threshold values are 3, 5 and 10.
- The **cell dominance rule** (also called the cell concentration rule) which applies to cells where a small number of data providers contribute a large percentage to the cell total.

Setting values for these rules is the responsibility of individual data custodians. The value used will depend on that custodian's assessment of the identification risk. Some datasets may be more sensitive than others while legislation and organisational policies will also influence the values that are set and applied.

Frequency rule

The frequency rule is best applied to count tables. While it can be used for magnitude tables, the underlying frequency data (i.e. the number of contributing units for each cell) is needed to apply the rule. The frequency rule is illustrated in example 1.

Sometimes zero cells (cells with no contributors, or cells where all respondents reported a zero value for a magnitude table), or 100% cells (where all contributors within a category have a particular characteristic) are also considered confidential.

A 100% cell that may not require protection is included in example 1. In this example all 15–19 year olds have a low income (20). This result may not be unexpected or sensitive and the cell may not require protection. However, if this same example related to a very small community and the 100% cell was instead for 15–19 year olds with high income then the cell may need to be protected from an identification risk as the result is unexpected and may be considered sensitive.

Cell dominance rule

The cell dominance rule is also referred to as the (n,k) rule and is used for magnitude tables.

According to the cell dominance rule, a cell is regarded as unsafe if the combined contributions of the 'n' largest members of the cell represent more than 'k'% of the total value of the cell.

The 'n' and 'k' values are determined by the data custodian. The cell dominance rule is illustrated in example 2.

Example 1: The frequency rule

Applying a frequency rule of 5, any cell with less than 5 contributors poses a disclosure risk. In the table below, the age group 50-59 years has 4 contributors in the low income cell and would be protected.

Age	Income			
	Low	Medium	High	Total
15-19	20	0	0	20
20-29	14	11	8	33
30-39	8	12	7	27
40-49	6	18	24	48
50-59	**4**	5	14	23
60+	12	9	7	28
Total	64	55	60	179

Example 2: The cell dominance rule

Consider a magnitude table of total profit by industry. Within the table, there may be one cell that presents total profit for the widget industry of $300m. To apply the cell dominance rule to this cell, the profit of all contributors to the cell must be known.

Say there are eight contributors to that cell, denoted by brand A through H (shown in the table at the right).

Using a (2,75) rule, the two largest contributors (brands A and B) must not contribute more than 75% of the cell total (that is, the total widget industry profit).

The combined profit of the two largest contributors (150+93=243) is 81% of the total widget industry profit (300), which is greater than 75% of the total.

Therefore, using the (2,75) rule the profit made by the widget industry is unsafe and should be protected.

Widget brand	Profit ($m)
A	**150**
B	**93**
C	21
D	13
E	8
F	8
G	6
H	1
Total profit for widget industry	**300**

For more information about confidentiality techniques or to provide feedback on this series, please email: inquiries@nss.gov.au

Confidentiality – how to confidentialise data: the basic principles

Techniques to confidentialise data

There are various techniques used to confidentialise data that pose an identification risk – either as an unsafe cell in aggregated data, or as unsafe records in microdata. These are grouped into two broad groups: data reduction methods and data modification methods.

Data reduction methods

These methods maintain confidentiality of respondents by selecting appropriate aggregations or in presentation of data.

1. Combining (or collapsing) categories

This method involves combining several response categories into one, or reducing the amount of classificatory detail available in a table or in microdata. It is often used for magnitude data and sometimes for count data.

Combining or collapsing categories is best used when a handful of responses have a small number of contributors, or when a table is very detailed and there are many small cells.

A good knowledge of the subject is important when combining or collapsing categories to ensure that the new category groupings are relevant to data users.

2. Primary and secondary suppression

Data suppression involves not releasing information for unsafe cells, or, if nothing else works, deleting individual records or data items from the microdata file.

If a table contains totals, it may be possible to calculate the value of a suppressed cell by subtracting the value of other cells from the total. At least one additional cell may also need to be suppressed to prevent identification.

A cell that is suppressed because it fails one of the confidentiality rules is called a **primary** suppression cell. The suppression of other cells (to prevent the disclosure of a primary suppression cell) is called **secondary** suppression or consequential suppression.

Example 3: Combining or collapsing categories

Classifications which are very detailed, such as geography, country of birth, industry or occupation, can be collapsed or combined to a broader level. For example, occupation categories for nurses and doctors may be combined into a 'medical profession' category.

For quantitative and/or continuous data items such as income or age, categories can also be combined in ranges. For example, age is often expressed in 5 year groups.

Top and bottom coding is a particular type of data reduction method that combines or collapses categories according to upper or lower thresholds. Values above an upper limit, or below a lower limit, are placed in an open-ended range. For example, all people over the age of 80 may be given an age value of 80+.

Example 4: Primary and subsequent suppression

In example 1, the table cell containing the number of low income earners aged 50-59 was identified as an unsafe cell using a frequency 'rule of 5'. This cell could be protected by suppressing it, as shown by the 'X' below.

However, it would still be possible to work out the value of the cell by using the remaining values. For example, low income earners aged 50-59 could be derived by subtracting the medium and high income earners from the total (23 minus 5 minus 14). Subsequent suppression is needed to protect the unsafe cell. Ways to do this include secondary suppression and concealing totals.

Secondary suppression

Age	Income			
	Low	Medium	High	Total
15-19	20	0	0	20
20-29	14	11	8	33
30-39	8	12	7	27
40-49	~~6~~ **Y**	~~18~~ **Y**	24	48
50-59	**X**	~~5~~ **Y**	14	23
60+	12	9	7	28
Total	64	55	60	179

In this table, the cells with the smallest values have been suppressed. The cell with an X indicates the unsafe cell (the primary suppression). The cells marked Y would also be suppressed to prevent the unsafe cell being worked out indirectly from the totals (secondary suppressions).

Concealing totals

Age	Income			
	Low	Medium	High	Total
15-19	20	0	0	20
20-29	14	11	8	33
30-39	8	12	7	27
40-49	6	18	24	48
50-59	**X**	5	14	~~23~~**>19**
60+	12	9	7	28
Total	~~64~~ **>60**	55	60	~~179~~ **>175**

Using this method, the cell with an X would be suppressed and subsequent row and column totals would be concealed, as shown above, to prevent the unsafe cell being worked out.

Data modification methods (perturbation)

These methods maintain respondent confidentiality by altering the identifiable data in a small way without affecting aggregate results.

Data rounding

Data rounding involves slightly altering small cells in a table to ensure results from analysis based on the data are not significantly affected, but the original values cannot be known with certainty. Data rounding may be random or controlled.

1. Random rounding is used in count tables and involves replacing small values that would appear in a table with other small random numbers. This results in some data distortion so that the sum of cell values within or between tables will not equal the table total.

Random rounding to base *X* involves randomly changing every number in a table to a multiple of *X*. For example, random rounding to base 3 (RR3) means that all values are rounded to the nearest multiple of 3. Each value, including the totals, is rounded independently. Values which are already a multiple of 3 are left unchanged.

2. Graduated random rounding is a rounding method used for magnitude tables. It is similar to random rounding, however after specified cell sizes the rounding base increases. That is, a small number will have a smaller rounding base than a large number. This ensures the protection offered does not diminish for large-valued cells.

3. Controlled rounding is a form of random rounding but it is constrained to have the sum of the cells equal to the appropriate row or column totals within a table. This method may not provide consistency between tables.

Sampling

Sampling provides some protection against identification risks because it reduces the certainty about whether a particular individual or organisation is in the data. A value may be unique in the sample, but not necessarily unique in the population.

Example 5: Random rounding to base 3

Before rounding

Age	Income			
	Low	Medium	High	Total
15-19	20	0	0	20
20-29	14	11	8	33
30-39	8	12	7	27
40-49	6	18	24	48
50-59	4	5	14	23
60+	12	9	7	28
Total	64	55	60	179

After rounding

Age	Income			
	Low	Medium	High	Total
15-19	21	0	0	21
20-29	15	12	9	33
30-39	9	12	6	27
40-49	6	18	24	48
50-59	3	6	15	24
60+	12	9	6	27
Total	63	54	60	180